Thesis/
Reports
Fernandez,
C. A.

A Structural Based Prediction System for Modelling Tree Mortality

C -JV-11222063-168

A Structural Based Prediction System for Modelling Tree Mortality

Clara Antón Fernández and Robert E. Froese

Prepared for

Moscow Forest Sciences Laboratory
Rocky Mountain Research Station
USDA Forest Service
JV-11222063-168

School of Forest Resources and Environmental Science
Michigan Technological University

RMRS Library
USDA Forest Service
APR 25 2007
240 W Prospect Rd
Fort Collins CO 80526

December 22, 2006

Executive Summary

We explore methods for implementing errors-in-variables (EIV) techniques for a logistic equation used to predict tree mortality. EIV methods account for bias that may occur in regression models when predictor variables are subject to error, such as basal area per acre estimated from sample plots. Here, we investigate methods for EIV for non-linear models, and develop an EIV version of the individual-tree mortality equation used in the Inland Empire variant of FVS. We use a limited set of data from the Permanent Plot program in Region 1 because a larger data set from the Forest Inventory and Analysis program was found to be not suitable. The EIV methodology we selected was Regression Calibration (RC) because it is straightforward, well accepted, and not obviously inferior to alternatives. RC-developed mortality equations differed meaningfully from equations developed ignoring sampling error, and in ways that conformed to expectations. These results were promising in initial evaluation through sensitivity analysis, but we were not able to answer the central question about potential bias because we lacked evaluation data. Future work should involve obtaining a larger calibration data set as well as independent validation data, and explore the effect of including an EIV mortality model in the larger FVS framework.

Table of Contents

EXECUTIVE SUMMARY I

TABLE OF CONTENTS II

TABLE OF FIGURES III

INTRODUCTION 1

OBJECTIVES 4

TERMS OF REFERENCE 5

MATERIALS AND METHODS 6

RESULTS 15

DISCUSSION 16

CONCLUSION AND RECOMMENDATIONS 19

LITERATURE CITED 20

Table of Figures

Figure 1. Proportion of alive/dead trees by species. Bar width is proportional to the number of trees in the data set. 10

Figure 2. Distribution of the four most important variables. 11

Figure 3. PBAL frequency distribution for western hemlock before (left) and after (right) regression calibration. 12

Figure 4. Western larch (top) and western hemlock (bottom) sensitivity graphs. 13

Figure 5. Sensitivity graphs for lodgepole pine. 14

Figure 6. Sensitivity graph for white pine. 14

Table of Tables

Table 1. Equations table. 15

Introduction

Forest growth and yield models are important tools in forestry, and the Forest Vegetation Simulator (FVS) is one of the most well-known and extensively used simulation models in North America. As our knowledge and available tools develop, improvements to FVS have been undertaken. Recently, the Prognosis individual tree basal increment model was revised to address the effect of sampling errors, or "Errors-In-Variables", on model parameterization and application. Stage and Wykoff (1998) called the revised methodology "Structural Based Prediction" (SBP). This and other revisions have proven useful, and are justifiably extended to other model sub-components; in particular, to the individual tree mortality model. The problem addressed in this research is to refine and apply methodologies to do so, given the different nature of the statistical model used to represent mortality and the results of recent work.

SBP is an alternative method for parameterizing and applying regression models, designed to reduce the bias in model predictions that can be present when predictor variables are subject to sampling error. While SBP appeared to be a useful advance, it was limited to one model sub-component, the basal area increment model. Importantly, the rationale for applying SBP applies equally to the mortality model as to the basal area increment model. Both models make use of predictor variables that are subject to sampling error, such as stand or plot basal area (SBA or PBA) or basal area in larger trees (BAL), potentially resulting in biased estimates of regression parameters. However, the basal area increment model is represented by a linear equation, while the mortality model is represented by a logistic equation, and SBP, as described by Stage and Wykoff (1998) is only applicable to the linear case.

Modeling Tree Mortality

Mortality is a discrete event. Tree status can take only one of two values: alive (0) or dead (1). There are few methodologies to model this type of behavior. The most common in tree mortality models is logistic regression, but in other fields dealing with mortality, such as medical science, survival analysis is more widely used. It has been argued that survival analysis cannot be applied successfully to individual tree mortality models due to the lack of age data in most inventories (Flewelling and Monserud 2002). Woodall et al. (2005) suggested using dbh (diameter at breast height) and αdbh (increment of dbh) instead of age as the measurement unit to indicate the change in an individual's status. Though this approach might lead to an interesting improvement in mortality modeling in the future, no advantages compared to the

traditional way to model individual tree mortality are evident so far. Therefore we take the classical approach to model the probability of mortality, the logistic equation:

$$P_m = 1 - \left[1 - \frac{1}{\left(1 + e^{\hat{a}X}\right)}\right]^t$$

Where P_m, is probability of mortality, βX is a linear combination of the regression coefficients and the predictors, and t accounts for the different mortality observation period lengths (Monserud 1976), in years. Mortality is usually observed by re-visiting permanent sample plots periodically and updating tree status.

Choosing the logistic form to model mortality has some advantages, the most obvious being that it implicitly bounds values between 0 and 1. Another important advantage is that the logistic model is by itself a direct probability model, and therefore, its only assumptions relate to the form of the regression equation. The assumptions for the logistic model are linearity of the continuous predictors with the log odds of the outcome and lack of interaction between the prediction variables (Harrell 2001). If the assumption of linearity is not met, the coefficients of the model will be underestimated and the model will lack power. This will result in Type II errors, and failing to detect actual functional relationships.

Errors-In-Variables

Structural Based Prediction

Stage and Wykoff's (1998) SBP is based upon a method of moments estimator presented by Fuller (1987) and extended explicitly to simulation models. SBP has two phases: estimation and application. In the estimation phase, the error in prediction variables is estimated and explicitly accounted for in parameter estimation, closely following Fuller's theory. In application and to generate predictions, the coefficients of the regression equation are dynamically modified, case-by-case, using current estimates of error in the simulated system. For sampling error in stand basal area (SBA) Stage and Wykoff used the usual sample-based estimator where multiple plots were established in a given stand. For basal area in the vicinity of an individual tree (PBA) they derived an estimator assuming a spatially random (Poisson) tree distribution. The authors demonstrated the technique using the Prognosis (Stage 1973) basal area increment model (Wykoff 1990), presented a sensitivity analysis of many assumptions, and included a case study

of the impact on larger model behavior when the SBP version is employed, albeit with artificial data.

Regression Calibration

Fuller's (1987) method-of-moments estimator and Stage and Wykoff's (1998) SBP technique are restricted to linear models and therefore, they are not applicable to the logistic case. For nonlinear models, the most recent literature describes five main approaches for accounting for measurement error in modelling. They include: regression calibration (Carroll and Stefanski 1990), simulation extrapolation (Cook and Stefanski 1994, Stefanski and Cook 1995), score function methods (Stefanski and Carroll 1987, Stefanski 1989, Nakamura 1990), likelihood and quasilikelihood (Carroll et al. 1984, Wang et al. 1996), and Bayesian methods (Gustafson 2004). The first two approaches are widely applicable, general methodologies, although they result in estimators that are only approximately consistent in general. The last four approaches share the advantage of resulting in estimators that are fully consistent estimators more generally than the first two approaches.

Regression calibration (Pierce and Kellerer 2004) and simulation extrapolation (SIMEX) are both simple and potentially applicable to any regression model, and they share the advantage that once the variable measured with error is replaced by the corrected variable, the data can be modelled as if the error does not exist. In other words, a model can be fitted with the usual, or preferred method. This is a very desirable characteristic since it implies that the model is parameterized with the corrected variable and the solution is universal. In application for prediction the error-prone variable has to be corrected for measurement error prior to running the model. This is in contrast to Stage and Wykoff's approach, where the observed values of the predictor variables are used and it is the regression parameters that are adjusted for each prediction. Additionally, no data from the fitting dataset have to be stored to correct for measurement errors in a new application. The main difference, in practice, between RC and SIMEX is that SIMEX is more computationally intensive. Therefore, regression calibration was the selected approach.

Regression calibration has become one of the most popular methods for adjusting for measurement error, particularly in logistic regression. We have to point out that, although this method is well established in logistic regression, effective and fairly well investigated (Pierce and Kellerer 2004), RC is an approximate correction method, it reduces the bias, but it does not necessarily produce consistent estimators.

The key idea of RC is predict the unobservable predictor **X** using the correspondent surrogates **W** (observation related with **X**) and **Z** (error free variables). Applied to individual tree mortality models this means that we use the measured PBA, basal area of larger trees in the vicinity of the trees (BAL) or SBA and the rest of the assumed error-free variables (diameter at breast height, crown ratio, etc.) to predict the correspondent real value of PBA, BAL or SBA.

The best linear approximation when α_{uu} (covariance matrix of measurement error or sampling variance) is estimated via external data, via internal replicates or when an external estimate α_{uu} is available was derived independently by Carroll and Stefanski (1990) and Gleser (1990). This approximation assumes the classical additive error model (**W** = **X** + **U**, where **U** is the observation error) where conditional on (**Z**, **X**) the errors have mean zero and covariance matrix α_{uu}.

The algorithm of the best linear approximation when an external estimate α_{uu} is available is:

$$E(\mathbf{X}_i|\mathbf{Z}_i, \overline{\mathbf{W}}_i) \approx \hat{\Gamma}_w + \left(\hat{\approx}_{xx}, \hat{\approx}_{xz}\right) \begin{matrix} \hat{\approx} \\ \approx \end{matrix} \begin{matrix} \hat{\approx}_{xx} + \hat{\approx}_{uu} & \hat{\approx}_{xz} \\ \hat{\approx}_{xz}^t & \hat{\approx}_{zz} \end{matrix} \approx \begin{matrix} \approx^{-1} \\ \approx \end{matrix} \begin{matrix} \overline{\mathbf{W}}_{i\bullet} \approx \hat{\mu}_w \approx \\ \mathbf{Z}_i \approx \overline{\mathbf{Z}}_{\bullet} \approx \end{matrix}$$

Equation 1. Best linear approximation algorithm

Where α in general denotes a covariance matrix for two random variables indicated by subscript, and μ_W indicates the mean of **W**. Note that we assume a linear relationship between **X** and (**W**, **Z**).

We adopt the same procedures for estimating the sampling variance (α_{uu}) as developed by Stage and Wykoff (1998).

Objectives

This research has the overall goal of re-engineering the Prognosis mortality model at an operational scale, using Structural Based Prediction. This goal is divided into four key objectives:

1. Develop and document a procedure for SBP that is statistically defensible for the logistic case, the latter being the formulation used for the mortality model;

2. Develop, fit and document a new formulation for the mortality model; and,

3. Evaluate the contribution of SBP by evaluating the new model using independent data.

Terms of Reference

This project is a component of a larger effort to redevelop the Prognosis mortality equation to include both SBP methods and examine the utility of lithology (soil parent material) as a predictor variable. Support for the larger effort includes a grant from the Macintire-Stennis program to Robert Froese as well as direct support from Michigan Technological University. The principal investigator is Robert Froese, who has been responsible for project coordination and contributed to methods development, analysis and reporting. Ms. Clara Antón Fernández is a doctoral student at Michigan Tech and has had primary responsibility for project activity and reporting.

Materials and Methods

Available Data

Two geographically extensive data sets were available from Froese's (2003) dissertation research. The larger set was obtained from the USDA Forest Service, Forest Inventory and Analysis program. This data set covers the entire geographic region to which the Northern Idaho variant of FVS is intended to apply. It includes nearly 65,000 individual tree measurements across 2,756 field locations, of which 2,697 or nearly 4% of tree records represent individual occurrences of mortality. Mortality was, however, assessed retrospectively; a "recent mortality" tree was a dead tree determined by the field crew to have died in the past five years, based on some established criteria. This data set was intended to form the fitting data set upon which the mortality model will be constructed, and this additional use leverages the original investment in obtaining it.

The smaller set was obtained from the USDA Forest Service Region 1 Permanent Plot Program, known as the "monitoring" data. This set includes an arbitrary sample of treated, regenerating stands in the Rocky Mountain Region. The survey design includes a control and several treated clusters of plots where measurements are made on individual tree attributes, including species, diameter, and height. Froese's (2003) monitoring data included data deemed of adequate state for publication by the Forest Service as of December 2002, which included areas on Idaho Panhandle (Kanisku, Coeur d'Alene and St. Joe), Flathead, Kootenai and Lolo National Forests. A total of 34,243 tree measurements across 189 stands were available that had two observations in time and were alive at re-measurement 0 (establishment). Mortality was observed at re-measurement 2 in 1,504 (4,39%) of the trees. This data set was intended to be reserved for validation.

Regression Calibration

Regression calibration algorithm (Equation 1) is well defined in Carroll and Stefanski (2006) but it is only currently implemented in software for the SAS statistical package (SAS Statistical Institute, Cary, NC). Since all statistical analyses were conducted using R software (R Development Core Team 2006), we coded the RC algorithm in R.

Model Development

Finding the best logistic model formulation is not an easy task. Linearity, interaction and significance have to be taken into account. We have followed roughly the methodology suggested by Hosmer and Lemeshow (2000). The first step is a univariable analysis of each variable. A test for significance of each predictor variable was performed using a likelihood ratio test with 1 degree of freedom. We selected the variables with a p-value < 0.25. Only the variables selected in the first step are used for step 2.

The second step consists of testing the linearity assumption, which is a circular process. We start by checking the linearity of one of the variables, and transform it if necessary in order to meet the linearity assumption. We then check the next variable, and so on. When we have finished checking for linearity and transforming the last variable it is necessary to go back to the first variable and check if it is still linear in the logit. This process is repeated until the solution is stable.

The next step is checking the ratio likelihood test p-value, removing all variables that have a p-value larger than 0.05. Afterward, we check for interaction, adding to the model a term that is the product of the transformed variables we want to test for interaction. After all the necessary terms have been added, we check again for linearity in all the continuous variables of the model. This process will stop when all the results are consistent. The last step of the model fitting is adding the variables that were excluded from the analysis in the first step and testing their statistical significance. This last step is intended to detect variables that were not significantly related to the response variable by themselves but in the presence of the other transformed variables make an important contribution. If any of the variables added in this last step are significant the model is tested again from step 2 with the new variable/s included.

Linearity

There are several methods that have been used to test linearity. We chose cubic spline function (Devlin and Weeks 1986) because is an appropriate method and it is easily implemented (Harrell and Lee 1985, Harrell 2001). This method consists of adding to the model the constructed spline lines as variables, running the model and testing their coefficients. The null hypothesis in this case is $\alpha_i = \ldots = \alpha_j = 0$ where $\alpha_i \ldots \alpha_j$ correspond to the coefficients of the constructed spline lines.

The cubic spline function method is an adequate method to test linearity of the continuous predictors with the log odds of the outcome. However, the cubic spline method does not suggest simple transformations of the variable. We chose fractional polynomials (Royston and Altman 1994, Sauerbrei and Royston 1999, 2002) to help finding the best simple transformations since this is one of the most commonly used methodologies.

Model Evaluation

We use the unweighted sum of squares test of Copas (1989) as modified by le Cessie and van Houwelingen (1991) to test the goodness-of-fit for logistic regression. This test does not require one to construct intervals of predicted probability, is simple and is as good as any alternative (Hosmer et al. 1997).

The Copas-le Cessie-van Houwelingen test is a one degree of freedom omnibus test that generally has more power than the Hosmer-Lemeshow test (Hosmer and Lemeshow 2000), although, like all omnibus tests, it has low power against directed alternatives. In this test values above the α-level indicate no lack of fit.

Results

Suitability of Available Data

Originally, two datasets were available: the FIA dataset and the monitoring dataset. We intended the first to be used to fit the model, and the second one to validate the model. The preliminary mortality model fitted using FIA data revealed some unusual patterns, however. Further analysis suggested that these unexpected patterns were due to the unreliability of the estimates of the date of death from temporary plots, as already noted by Flewelling and Monserud (2002). Therefore, we made the monitoring dataset the principal fitting dataset, and rely here upon fit statistics for model evaluation and validation.

Mortality in the monitoring data is assessed by remeasured tree survival data, and therefore more reliable. One of the benefits from using the monitoring data is that crown ratio, an important predictor in other mortality models, was available. On the other hand, the monitoring dataset is much smaller than the FIA one and this has two drawbacks: the first is that some of the species have not enough trees to fit a model, the second is that we don't have any more independent data to validate the model. Since it has not been possible to obtain new independent data, the effects of applying regression calibration have been evaluated with sensitivity analysis. Another drawback of the monitoring dataset is that it is basically formed by young stands.

Model Development

The original dataset included 10 species of softwood trees and the hardwoods group, totaling 34,243 trees. However, not all the species were equally represented in the dataset (Figure 1). Western redcedar and subalpine fir had only 15 and 29 mortality events observed, respectively. For this reason both species were excluded from the analysis. For grand fir a statistically significant model could not be fit. Hence, we have been able to fit mortality models for 8 species: western larch, lodgepole pine, Engelmann spruce, white pine, ponderosa pine, Douglas-fir, western hemlock and the hardwoods group.

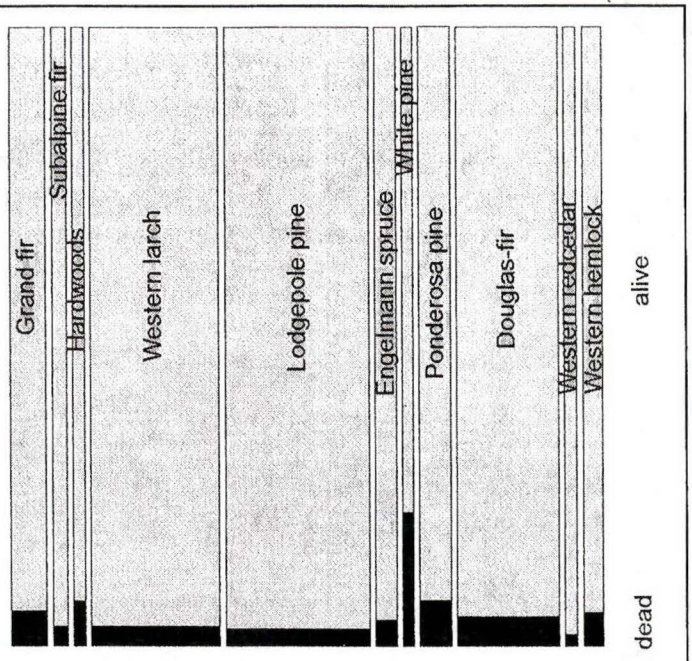

Figure 1. Proportion of alive/dead trees by species. Bar width is proportional to the number of trees in the data set.

The initial set of variables studied included diameter at breast height (DBH), point basal area (PBA), point basal area of larger trees (PBAL), stand basal area (SBA), crown ratio (CR), and a dummy variable to distinguished between managed and unmanaged stands (1-managed, 0-unmanaged). From the original set of variables only SBA is not significant in any of the models.

The distribution of the rest of the variables among their range is poor, except in the case of crown ratio. PBAL and DBH are especially poorly distributed with most of the data in a very reduced range (Figure 2).

For every species we fitted two models. We fitted one model ignoring the sampling error, which we call the naïve approach to the problem. The second model was fit correcting for the sampling error with regression calibration, which we call the RC approach.

The variables that emerged as statistically meaningful depended on the approach. In the naïve approach DBH appears only in 2 of the models, while it is not significant in any of the models in the regression calibration approach. In the naïve approach CR and PBAL are part of every model except in the case of ponderosa pine, where is replaced by PBA. In the regression calibration approach CR is a significant predictor for every species except for white pine, while PBAL is a good predictor except for hardwoods, and Engelmann spruce. The mortality equations fitted for both approaches and for all the species are described in Table 1.

In earlier versions of the models DBH was statistically significant for western larch and lodgepole pine. However, when plotted, the slope was small, and the trend of the mortality response along the DBH range was not biologically consistent. Therefore, we decided not to include the variable in the models for these two species.

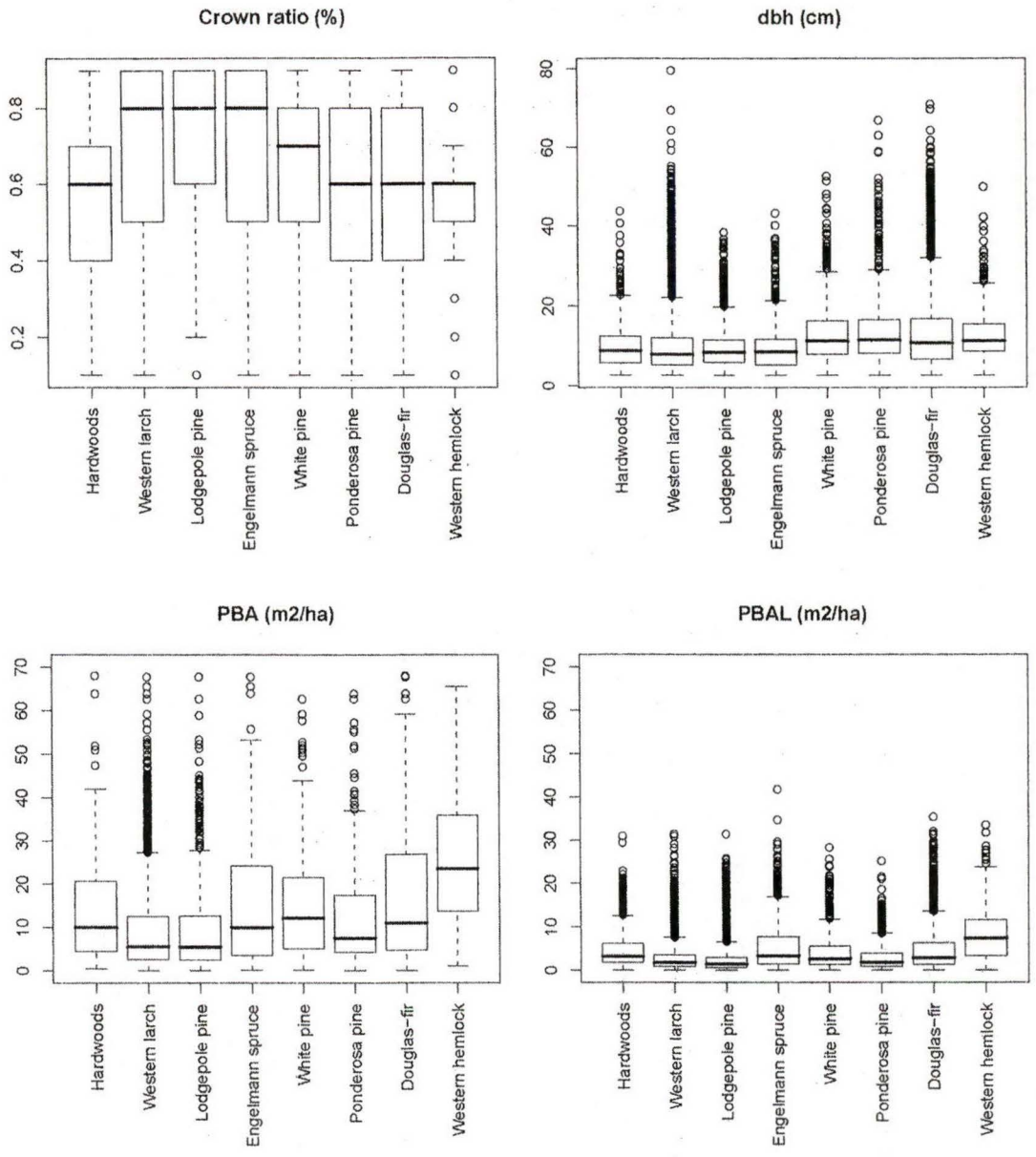

Figure 2. Distribution of the four most important variables.

During the process of applying regression calibration we have noticed that the results differ depending on which variables were included, and if the variables included are or not transformed. The best option has not been investigated. We have chosen the untransformed error-free variables that are usually included in mortality models, namely DBH, CR and the dummy variable managed.

The main effect of RC on the variable with sampling error is the shrinkage of the distribution of the variable towards its mean (Figure 1). As a result of applying regression calibration a small proportion of the data (PBA or PBAL) became negative. We decided to restrict them to zero. These negative values correspond to high error variances.

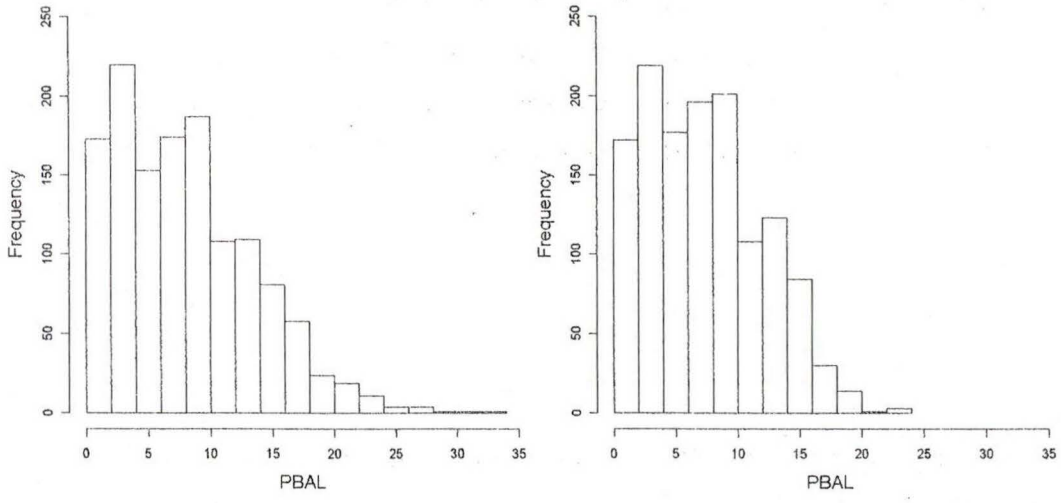

Figure 3. PBAL frequency distribution for western hemlock before (left) and after (right) regression calibration.

Fitted Model Performance

In the naïve approach trees in managed stands have, in general, higher probability of mortality than in unmanaged stands. This difference tends to decrease in the RC approach (Table 1 and Figure 5).

The residual deviance under the regression calibration approach is always higher than under the naïve approach, except for western hemlock. However, this increase in residual deviance is not necessarily accompanied by a decrease in the overall goodness-of-fit (Table 1).

With respect to PBAL, in western larch and western hemlock the mortality is similar in both models until PBAL reaches 10 m^2/ha, when the RC-model gives clearly more probability of mortality for the same PBAL. In western larch the main difference in CR occurs in the smaller values of CR between both approaches (Figure 4).

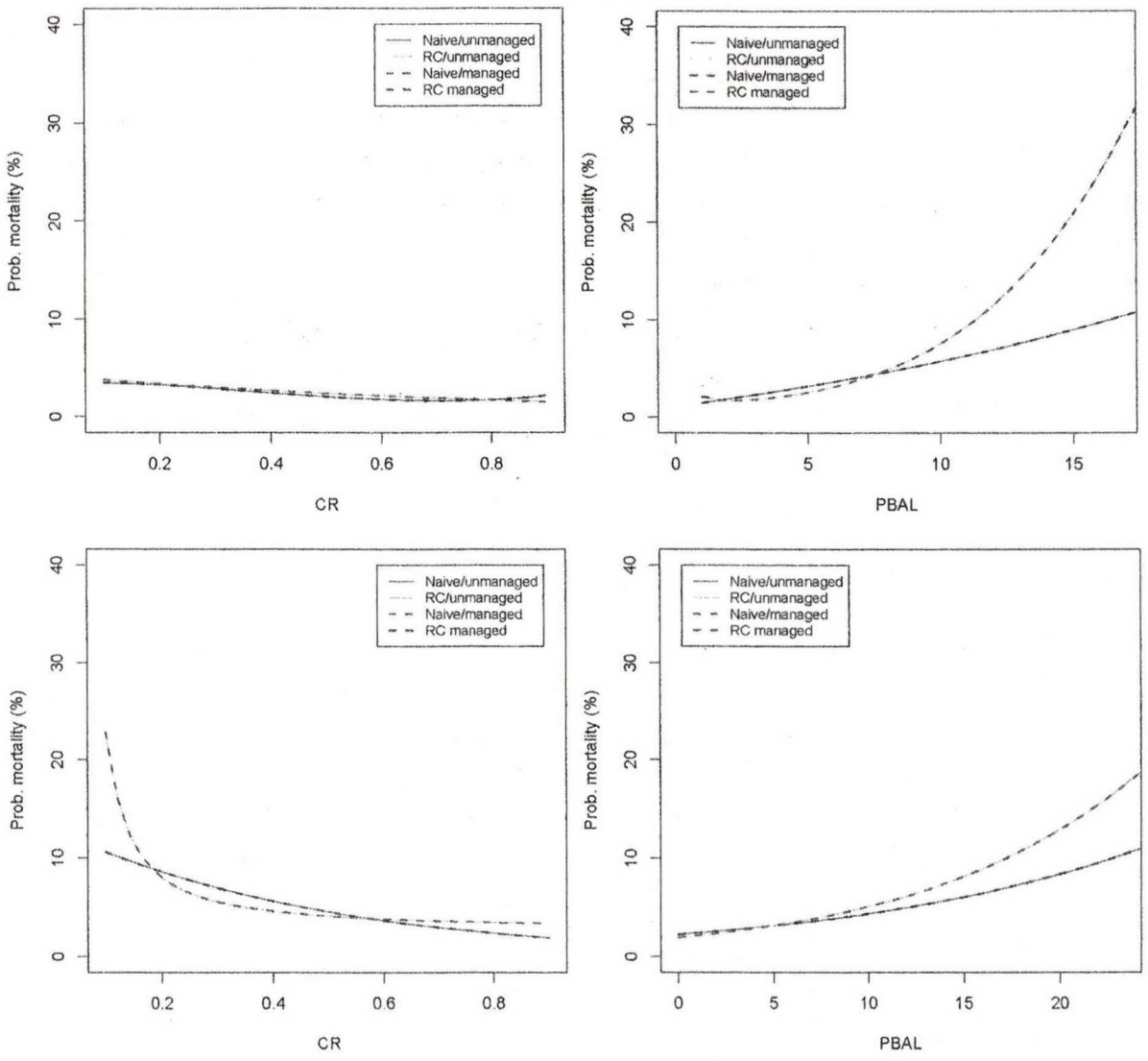

Figure 4. Western larch (top) and western hemlock (bottom) sensitivity graphs.

The sensitivity graphs for lodgepole pine (Figure 5) show a decrease in the slope of the PBAL curve for the RC approach and an increase in the CR slope for the RC approach versus the naïve approach.

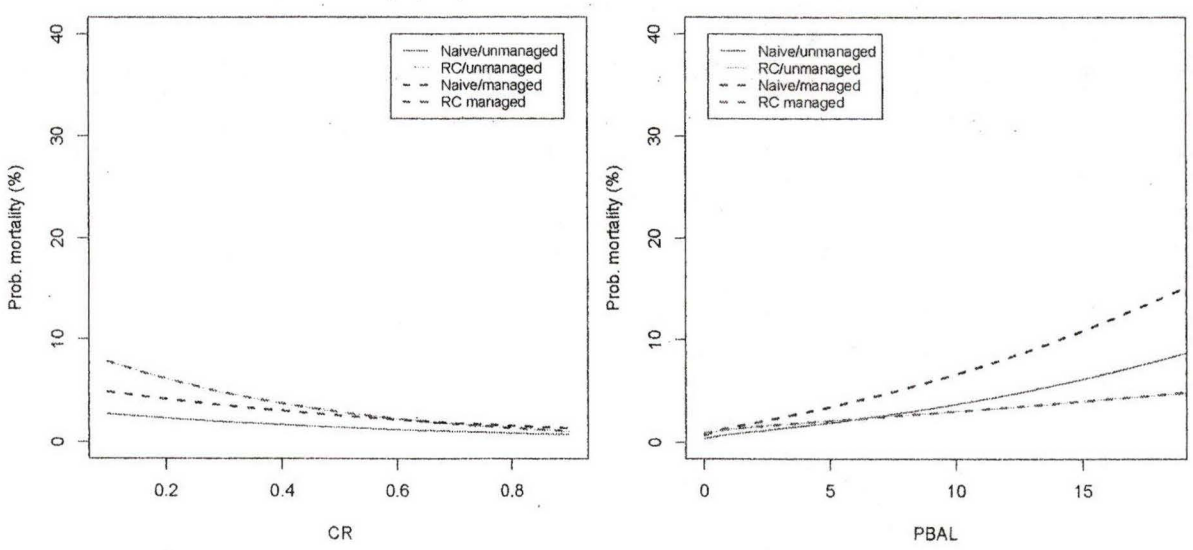

Figure 5. Sensitivity graphs for lodgepole pine.

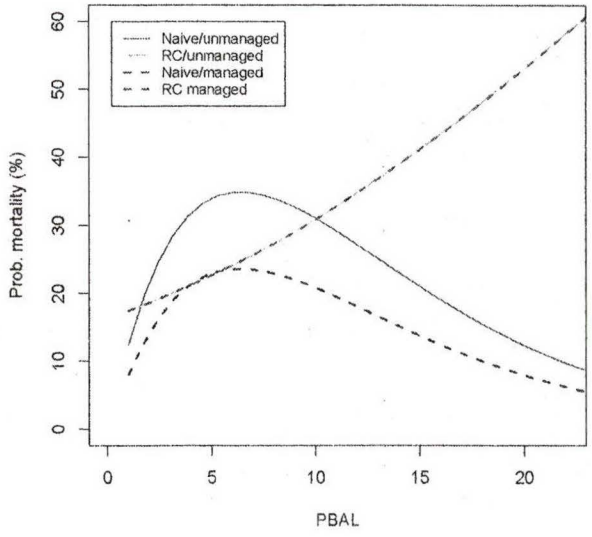

Figure 6. Sensitivity graph for white pine.

Lodgepole pine (Figure 6) has an unexpected shape in the naïve approach.

	NAIVE	Residual Deviance	gof	REGRESSION CALIBRATION	Residual Deviance	gof
Hardwoods	$-4.43904673-2.45772020\cdot CR+0.09842712\cdot PBAL$	350.7	0.49	$-5.9481650-1.7546127\cdot\log(CR)$ $-0.6922406\cdot managed$	354.7	0.889
Western larch	$-6.5010374+6.4327017\cdot CR^3\cdot\log(CR)+$ $+0.6647012\cdot\sqrt{PBAL}$	2141.2	0.936	$-7.898635-1.171606\cdot CR$ $+2.659154\cdot\sqrt{PBAL+1}$ $-1.921760\cdot\log(PBAL+1)$	2161.7	0.883
Lodgepole pine	$-6.5760569-1.6185786\cdot CR+0.7289831\cdot\sqrt{PBAL}$ $+0.5941602\cdot managed$	2135.8	0.744	$-5.0211160-2.5324038\cdot CR$ $+0.4024486\cdot\sqrt{PBAL}$	2179.4	0.998
Engelmann spruce	$-4.32162137\pm2.80052468\cdot CR+0.07124379\cdot PBAL$	433.3	0.193	$-3.009574-4.061720\cdot CR$	444.9	0.411
White pine	$-8.1067633.795323\cdot\sqrt{PBAL+1}$ $-0.984442\cdot\sqrt{PBAL+1}\cdot\log(PBAL+1)$ $-0.470651\cdot managed$	679.4	0.58	$-4.01790814+0.07371409\cdot PBAL$	690.6	0.979
Ponderosa pine	$-4.44747145+0.05063622\cdot(PBA+1)$ $-14.00237560\cdot\frac{1}{(PBA+1)^2}$ $-2.03681455\cdot CR-0.03479846\cdot DBH$ $+0.95360629\cdot managed$	957.4	0.714	$-6.1254572-1.8709418\cdot CR$ $+1.2545776\cdot\log(PBAL+1)$ $+0.9397457\cdot managed$	970.0	0.601
Douglas-fir	$-8.54679912+0.88157464\frac{1}{\sqrt{CR}}$ $+1.73767137\cdot\log(DBH)-0.44489887\cdot(\log(DBH))^2$ $+0.04393654\cdot PBAL+0.32948631\cdot managed$	2407.8	0.059	$-6.04910036+0.21990026\frac{1}{CR}$ $+0.04200316\cdot PBAL$	2448.3	0.962
Western hemlock	$-4.75728512-2.19435475\cdot CR+0.06721328\cdot PBAL$	526.6	0.132	$-6.59934316+0.22792923\frac{1}{CR}$ $+0.09730162\cdot PBAL$	523.0	0.951

Table 1. Equations table.

Discussion

The equations of the models in Table 1 may seem unnecessarily complicated, especially if compared with previous mortality equations (Hamilton 1986, Teck and Hilt 1990, Monserud and Sterba 1999). The transformations of the variables in the equations are necessary to ensure linearity in the logit of the predicted variable, a very important assumption in logistic regression (see Methods section). However, these transformations make it difficult to observe the consequences of applying regression calibration by simply looking at the coefficients. A good alternative is to compare the results of sensitivity analyses.

After applying regression calibration a small part of the data of the variable adjusted for sampling error became negative. This could be due to an excessive correction of the data with high error variances. These possible excess of correction might have affected other trees, though this possibility has not been investigated.

The unusual mortality response to PBAL for white pine (Figure 6) could have a number of reasons. This could be due to overfitting of the model (672 trees 146 dead). However, the high rate of mortality in white pine (Figure 1) suggests that the reason for the unusual shape of Figure 6 is that observed mortality in the dataset is not exclusively the background morality, but it is probably the combined effect of background mortality and a pest effect. We are mindful of the presence of white pine blister rust (*Cronartium ribicola*) in the study area.

Stage and Wykoff (1998) reported a higher variance under the SBP approach when compared to the naïve approach. In a similar way the RC approach has, for all models, a higher residual deviance than the naïve approach. Stage and Wykoff also reported an increase in R^2 in the SBP approach when compared to the naïve approach. Similarly, most of the models have a higher goodness-of-fit (gof) value. For those species where a decrease in the gof occurred, the differences are small.

Stage and Wykoff (1998) found that by applying the structural based prediction procedure the coefficient of the competition variables (PBAL, SBA) increases substantially in magnitude, while the crown ratio coefficient decreases. In our analysis the results do not show such a clear trend; they can even seem ambiguous at first look. However, contrary to the linear case, the effect of the sampling error in the multivariate logistic case can be under- or overestimate the effect of the variable (Greenland 1980, Kupper 1984). Moreover, unless the variables measured with

error are independent of the variable with sampling error, the coefficients of the variables measured without error are also biased in general (Carroll 1998). Hence, these effects are not surprising, although they have to be cautiously considered under the limited scope of the available data. It has also to be taken into account that the correlation between CR and competition might have affected the results. Trees in stands with low density, and dominant trees in dense stands tend to have relative high crown ratio. This can make the two factors fairly redundant except in recently disturbed stands. Thus, in the absence of disturbance, crown ratio reflects the relative competitive status of the tree.

Especially unusual, when compared to previous mortality models, is that DBH only appears as a significant predictor in 2 of the 8 naïve models and in none of the RC models. The low occurrence of DBH as a predictor in the naïve models is probably due to the small range of DBH present in the data (Figure 2), with most of the data in a 25 cm range. The drop of the variable in the 2 species in the RC models could also indicate a shift of emphasis towards vigor and competition variables (CR, PBA and PBAL).

The RC model for hardwoods and Engelmann spruce lost the PBAL variable. This might indicate an effect similar to lodgepole pine (Figure 5), but increased until the point that PBAL cannot be considered a useful predictor any more. It could also be due to the lower number of trees in the last two species, since a reduced number of trees makes it more difficult to find a small slope significant.

One of the keys to adequately applying RC is the correct estimation of Σ_{uu}. We have adopted the methodology described by Stage and Wykoff (1998). However, Stage and Wykoff already mentioned that a refinement of the variance model is needed. Moreover, Lappi (2005) critiques many aspects of Stage and Wykoff's methodology. Lappi suggests that since the local density around a tree affects tree growth, Stage and Wykoff's (1998) assumption that the measurement error is uncorrelated with the true value is not applicable for local density, and therefore the assumption of classical measurement error is incorrect. Lappi also notes that the assumption of Poisson locations of trees is not likely valid. However, Lappi does not suggest any solution to these problems and, even though the estimations of Σ_{uu} proposed by Stage and Wykoff might not be perfect, the results seem coherent. Furthermore, since there are no obvious better alternatives at this time we consider Stage and Wykoff's approach the best option.

Despite all the uncertainties, regression calibration does seem to have an effect on the variables sampled with error, and it does also seem to influence which variables are included in the optimum model. One of the main questions, namely if regression calibration reduces the bias due to sampling error in individual tree mortality models could not, regrettably, be answered due to lack of validation data.

The limited quantity and scope of the data has restricted the conclusions with respect on how RC affects the final model and the predictions derived from it. However, RC is simple and potentially applicable to any regression model. RC has the advantage that once the variables are corrected, one can use them almost as if they were the true unobserved value. These characteristics open a remarkable range of possibilities that makes worthy a future effort to improve and better understand the methodology proposed.

Conclusion and Recommendations

Despite limited data, RC showed promise. Many of the expected benefits seemed apparent in these very preliminary model calibrations. This is an encouraging result.

The next steps in the development of an operational model involve developing an expanded fitting and testing data set. The monitoring data used in this study were only a subset of those collected in Region 1; more data should now be available in a processed form. The availability of remeasured FIA data in the region, which would permit direct rather than error-prone retrospective observations of mortality, should be explored. Other data sets in the region likely exist, and should be identified.

More data would permit validation of the RC model. This should be done independently, of the mortality equation alone, but also with the mortality equation set into the larger context of the FVS model, similar to what Stage and Wykoff (1998) did. Comparing model behavior with an RC mortality model, and with an RC mortality model and Stage and Wykoff's SBP diameter increment model, would permit careful examination of the composite effect.

An errors-in-variables solution via RC should be compared to SPB, particularly in the context of the diameter increment model. Because RC corrects before models are fit or applied, it may offer a unified approach. The practical and theoretical tradeoffs between methods need to be examined and quantified.

Literature Cited

Carroll, R. J. 1998. Measurement error in epidemiologic studies. Pages 2491-2519 *in* P. Armitage and T. Colton, editors. Encyclopedia of biostatistics John Wiley & Sons, Ltd, Chichester, England.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. Measurement error in nonlinear models. Chapman and Hall/CRC.

Carroll, R. J., C. Spiegelman, K. K. Lan, K. T. Bailey, and R. D. Abbot. 1984. On errors-in-variables for binary regression models. Biometrika **71**:19-26.

Carroll, R. J., and L. A. Stefanski. 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. Journal of American Statistical Association **85**:652-663.

Cook, J. R., and L. A. Stefanski. 1994. Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association **89**:1314-1328.

Copas, J. B. 1989. Unweighted sum of squares test for proportions. Appl. Statist. **38**:71-80.

Devlin, T. F., and B. J. Weeks. 1986. Spline functions for logistic regression modeling. Pages 646-651 *in* S. I. Inc, editor. Proceedings of the Eleventh Annual SAS Users Group International Conference. SAS Institute Inc, Cary, NC.

Flewelling, J. W., and R. A. Monserud. 2002. Comparing methods for modelling tree mortality. Pages 168-177 *in* N. L. Crookston and R. N. Havis, editors. Second forest vegetation simulation conference. USDA-Forest-Service

Fort Collins, USA: Rocky Mountain Research Station., Fort Collins, Colorado, USA, 12-14 February 2002.

Froese, R. E. 2003. Re-engineering the prognosis basal area increment for the Inland Empire. University of Idaho, Moscow.

Fuller, W. A. 1987. Measurement Error Models, New York.

Gleser, L. J. 1990. Improvements of the naive approach to estimation in non-linear errors-in-variables regression models. *in* P. J. Brown and W. A. Fuller, editors. Statistical Analysis of Measurement Error Models and Application. American Mathematics Society, Providence.

Greenland, S. 1980. The effect of misclassification in the presence of covariates. American Journal of epidemiology **112**:564-569.

Gustafson, P. 2004. Measurement error and misclasification in statistics and epidemiology. CRC/Chapman & Hall.

Hamilton, D. A., Jr. 1986. A logistic model of mortality in thinned and unthinned mixed conifer stands of northern Idaho. Forest Science **32**:989-1000.

Harrell, F. E. 2001. Regression Modelling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer 2001, New York.

Harrell, F. E., and K. L. Lee. 1985. The practical value of logistic regression. Pages 1031-1036 *in* Proceedings of the Tenth Eleventh Annual SAS Users Group International Conference.

Hosmer, D. W., T. Hosmer, S. Le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine **16**:965-980.

Hosmer, D. W., and S. Lemeshow. 2000. Applied Logistic Regression.

Kupper, L. L. 1984. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. American Journal of epidemiology **120**:643-648.

Lappi, J. 2005. Plot size related measurement error bias in tree growth models. Canadian Journal of Forest Research **35**:1031-1040.

le Cessie, S., and J. C. van Houwelingen. 1991. A goodness of fit test for binary regression models, based on smoothing methods. Biometrics **47**:1267-1282.

Monserud, R. A. 1976. Simulation of forest tree mortality. Forest Science **22**:438-444.

Monserud, R. A., and H. Sterba. 1999. Modeling individual tree mortality for Austrian forest species. Forest Ecology and Management **113**:109-123.

Nakamura, T. 1990. Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. Biometrika **77**:127-137.

Pierce, D. A., and A. M. Kellerer. 2004. Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. Biometrika **91**:863-876.

R Development Core Team. 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Royston, P., and D. G. Altman. 1994. Regression using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Appl. Statist. **43**:429-467.

Sauerbrei, W., and P. Royston. 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J. R. Statist. Soc. A **162**:71-94.

Sauerbrei, W., and P. Royston. 2002. Corrigendum: Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J. R. Statist. Soc. A **165**:399-400.

Stage, A. R. 1973. Prognosis model for stand development. RP-INT-137, USDA For. Serv.

Stage, A. R., and W. R. Wykoff. 1998. Adapting distance-independent forest growth models to represent spatial variability: effects of sampling design on model coefficients. Forest Science **44**:224-238.

Stefanski, L. A. 1989. Unbiased estimation of a nonlinear function of a normal mean with appication to measurement error model. Communications in Statistics, Series A **18**:4335-4358.

Stefanski, L. A., and R. J. Carroll. 1987. Conditional scores and optimal scores in generalized linear measurement error models. Biometrika **74**:703-716.

Stefanski, L. A., and J. R. Cook. 1995. Simulation-extrapolation: The measurement error jackknife. Journal of the American Statistical Association **90**:1247-1256.

Teck, R. M., and D. E. Hilt. 1990. Individual-tree probability of survival model for the Northeastern United States. Research Paper Northeastern Forest Experiment Station NE-642, -USDA-Forest-Service.

Wang, N., R. J. Carroll, and K. Y. Liang. 1996. Quasilikelihood estimation in measurement error models with correlated replicates. Biometrics **52**:401-411.

Woodall, C. W., P. L. Grambsch, and W. Thomas. 2005. Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. Ecological Modelling **189**:199-208.

Wykoff, W. R. 1990. A basal increment model for individual conifers in the Northern Rocky mountains. Forest Science **36**:1077-1104.